图书情报工作
第 63 卷 第 22 期 2019 年 11 月

我国图书情报领域期刊论文的科学数据引用特征研究*

■ 丁文姚 李健 韩毅
西南大学计算机与信息科学学院 重庆 400715

摘要：［目的/意义］探索期刊论文科学数据引用特征与规律不仅有助于描述学科领域对科学数据的利用情况，还能够揭示学术成果表达中的数据引用模式。［方法/过程］以我国图书情报领域 6 种期刊 2017 年与 2018 年第一期刊载论文为样本，结合国家标准《信息技术 科学数据引用》的引用元素，采用内容分析法从 9 个维度对样本论文的科学数据引用行为进行数据编码，应用统计学方法描述图书情报领域期刊论文科学数据引用特征并探索不同维度特征间的关联关系。［结果/结论］图书情报领域期刊论文广泛引用来自国内外的统计整理类科学数据，对期刊论文中个人研究科学数据的引用量较大；科学数据引用标注方式与科学数据类型存在一定对应关系，但多样化的标注方式缺乏统一性；二手引用现象较为突出，二手引用程度与科学数据创建者类型相关。

关键词：图书情报领域 期刊论文 科学数据引用 引用特征
分类号：G250
DOI：10.13266/j.issn.0252-3116.2019.22.013

1 引言

伴随数据密集型科学研究范式的兴起，科学数据越来越成为驱动科学研究的重要力量，不仅自然科学研究依赖于数据的深入分析与挖掘，人文社会科学研究也有明显的数据化发展趋势，科研工作者在实践研究中的科学数据生成与引用已成为数据时代科研活动的普遍现象。

众多研究认为科学数据和学术论文同样重要[1]，科学数据引用在识别科学数据归属、促进科学数据共享、衡量科学数据影响力、推动数据计量研究等方面能够发挥重要作用，因而需要正确规范科学数据的引用实践；同时，科学数据的引用规范受到国内外相关机构组织的持续关注，2018 年国务院发布《科学数据管理方法》和国家标准《信息技术 科学数据引用》（GB/T 35294-2017）[2]充分说明我国对规范科学数据引用的重视。

科学数据引用是一种显著而普遍的信息行为方式[3]，分析科学数据引用行为不仅有助于描述学科领域对科学数据的利用情况，还能揭示科研工作者的科学数据引用行为特征与规律，从而为完善科学数据引用行为理论基础提供可靠依据。目前国内外关于科学数据引用行为研究已积累一定成果，相关研究主要基于内容分析方法，从论文的科学数据引用和科学数据库的被引两个视角展开。

（1）从论文的科学数据引用角度来看，相关研究侧重于从不同方面揭示不同学科领域期刊论文的科学数据引用特征。比如，丁楠等从引用频次和引用行为规范角度调研我国图书情报领域期刊和社会学领域期刊论文科学数据引用时发现，图书情报领域的规范引用比率呈现逐年提高，但非规范引用情况仍然明显[4]；社会学领域的科学数据引用行为普遍，各种年鉴及人口调查资料是主要引用数据来源，但不规范引用较多，而对已发表论文中数据的引用行为较为规范[5]；屈亚杰等[1]从创建者、类型、被引次数、访问方式、更新次数、规模及时间跨度等方面描述了美国校际社会科学数据共享联盟存储库（ICPSR）论文的科学数据分布特征；S. C. Williams[6]调研了农作物领域期刊论文引用的科学数据类型，发现该领域主要引用已发表文章中的数据集、数据补充文件以及气象站、数据仓储中的自然数据；H. Park 等[7]调研生物医学期刊论文的科学数

* 本文系四川省教育厅人文社会科学重点研究基地四川学术成果分析与应用研究中心资助项目“大数据环境下科学数据的引用行为研究”（项目编号：SCAA18-022）研究成果之一。
作者简介：丁文姚（ORCID：0000-0002-4789-2980），硕士研究生；李健（ORCID：0000-0002-2934-0872），副教授，博士，硕士生导师；韩毅（ORCID：0000-0001-7021-3229），教授，博士，博士生导师，通讯作者，E-mail：hanyi72@swu.edu.cn。
收稿日期：2019-01-18 修回日期：2019-05-12 本文起止页码：118-128 本文责任编辑：杜杏叶

据引用规范性发现,文章正文中的非规范引用形式比文章参考文献中的规范引用更为常见,这为数据引用评价造成了误差和困难。

(2)从科学数据库的被引角度来看,相关研究聚焦于从不同维度调研某类数据库的被引用分布情况。比如,赵蕊菡[8]从被引频次、年代、国别、引用位置调研 Earth System Science Data 中 178 篇数据论文的被引用分布情况,发现数据论文的重用现象较为普遍,但数据论文引用标准有待进一步完善;孟祥保等[9]调研了数据引文索引(DCI,Data Citation Index)收录的历史学、教育学、人口统计学、政府与法律、商业与经济领域的科学数据分布和引用率,发现高被引数据具有集中性,5 个学科约 90% 科学数据存在零被引现象,数据扩散广度和深度较为有限;R. S. Chen 等[10]调研 SEDAC(NASA Socioeconomic Data and Applications Center)中科学数据的被引用情况时发现,生态学和生物领域在生物多样性研究中对科学数据的引用最多,且存在大量跨学科数据的引用;T. Henderson 等[11]调研引用 CRAWDAD(Community Resource for Archiving Wireless Data At Dartmouth)科学数据的行为规范性时发现,有 11.5% 论文的引用没有采用该数据库提供的引用规范。

从已有典型科学数据引用成果可以看出,科学数据的引用以文本记录为表现载体,引用行为结果是了解科学数据引用特征较为直观的途径;科学数据引用行为研究主要通过内容分析探索不同学科领域的引用特征,研究维度包括科学数据的引用频次、科学数据类型、引用元数据元素、引用形式、引用位置、引用规范性等。然而,多数研究要么关注引用行为的规范性,要么描述各独立维度的科学数据类型分布特征,同时涉及科学数据内容和引用行为方式的研究较少。另外,尚没有学者关注不同维度引用特征之间的关联性。那么,科学数据类型与引用行为方式之间是否存在关联关系? 如果有,这些关联关系可能呈现出怎样的引用特征? 基于此,本研究以我国图书情报期刊论文的科学数据引用为对象,在描述引用特征的基础上聚焦于探究不同维度特征间的相关关系,旨在从科学数据类型和引用行为方式两方面揭示该领域的科学数据引用特征与特征间的关联关系。

2 数据与方法

2.1 概念内涵

(1)科学数据。《OECD 关于公共资助科学数据获取的原则和方针》[12]《科学数据共享工程技术标准研究报告》[13]《信息技术 科学数据引用》(GB/T 35294－2017)[2]等国内外相关规范对科学数据均有定义,但在表述上并不完全一致。

综合相关定义,科学数据应具有三方面特征:首先,在来源上,科学数据来源于客观实在的感知反映记录结果,是伴随科学研究活动过程产生的原始材料,具有科学性;其次,在内容和形式上,科学数据是反映人类社会科技活动或自然世界的客观事实记录材料,包括定性文字描述、定量数据等形式的事实原始记录和经加工整理的事实材料,内容具有原始素材性;最后,在功能效用上,科学数据不仅具有科学证据性,还具有可重用性与价值性,能为其他研究提供推理、讨论或计算基础。

基于此,本研究把科学数据定义为:为科学共同体认同的客观反映人类社会科技活动或自然界本质、特征、变化规律的证据性原始记录信息或经原始记录加工整理的数据材料。

科学数据的表现形式多种多样,从其表观形态大致可划分为非文字描述类与文字描述类。本研究聚焦于非文字描述类型科学数据,具体表现形式包括:定量数据、数据(集)、模型、图表、公式、算法、图片、音视频资料。

(2)科学数据引用。目前科学数据引用没有统一的概念描述。主要的定义描述如:研究人员在写论文时有必要在论文的参考书目或是页面的脚注部分引用所使用的资源,科学数据引用指通过一定的标识技术和机制描述科学数据资源、标识科学数据来源[3,14];指科研工作者引用科学数据作为论文观点的支撑数据,并以参考文献、脚注或文中注等方式,对其所引用数据提供数据参考的做法[15]。

总体上,相关定义的概念内涵大致相同,主要强调对引用的科学数据内容和来源出处进行参考标注。基于此,本研究将科学数据引用定义为:科研工作者以参考文献、脚注或文中注等方式对论文中引用的科学数据标注来源出处的信息行为。

2.2 数据来源

本研究的数据来源于 CNKI 数据库,数据获取的基本要求是:尽量覆盖较多期刊并保证样本论文质量。因此,根据 CNKI 数据库学科期刊导航提供的期刊影响因子数据,选择图书情报与数字图书馆领域期刊复合影响因子与综合影响因子(如表 1 所示)排序前六的期刊:《中国图书馆学报》《大学图书馆学报》《图书与情报》《图书情报知识》《国家图书馆学刊》《图书情报

工作》,所选 6 种刊物在一定程度上代表了本领域的研究水平;再以等距抽样法选取 6 类期刊 2017 年与 2018 年第一期刊载的论文为统计样本,去除其中编辑寄语、会议报导、简讯、回顾性散文等文献后共获得 175 篇样本论文。

表 1 样本期刊的复合影响因子与综合影响因子排名

期刊名称	复合影响因子	综合影响因子
中国图书馆学报	5.826	5.050
大学图书馆学报	2.445	2.255
图书与情报	2.341	2.090
图书情报知识	2.244	1.898
国家图书馆学刊	2.184	1.932
图书情报工作	2.140	1.871

注:数据来源 http://navi.cnki.net/knavi/Journal.html#

2.3 研究方法

(1)内容分析法。采用内容分析法对样本论文的科学数据引用内容与表现形式进行编码,并对编码结果进行描述统计分析。内容分析类目系统的编制依据是国家标准《信息技术 科学数据引用》(GB/T 35294 - 2017)的科学数据引用元素和研究目标需求,共从 9 个维度构建科学数据引用特征的内容分析类目,并通过前编码与后编码方式确定各类目的定类变量明细。其类目系统如表 2 所示:

表 2 内容分析类目与编码明细

分析类目	类目编码明细
期刊名称(X1)	中国图书馆学报,大学图书馆学报,图书与情报,图书情报知识,国家图书馆学刊,图书情报工作
期刊刊期(X2)	2017/01,2018/01
论文序号(X3)	自然整数
科学数据类型(X4)	原始记录类科学数据、统计整理类科学数据、研究成果类科学数据,不明
科学数据来源地区(X5)	中国,国外,不明
科学数据创建者类型(X6)	企业,政府行政机关,其他公共部门,社会团体、其他组织机构、个人研究、数据库、档案馆,科研机构,公共图书馆,学校,高校图书馆,不明
科学数据传播者类型(X7)	商业企业、社会团体组织、政府部门、互联网服务公司、科研机构、教育机构、其他互联网平台、期刊发表论文,专著、纸质资源、不明
科学数据引用呈现类型(X8)	直接引用图,基于引用数据作图,直接引用表,基于引用数据作表,公式,模型,文字论述,其他
科学数据引用方式(X9)	(文字类)标注文后参考文献,页脚注释标注文后参考文献,图表名称标注文后参考文献,段落正文说明来源并标注参考文献,图表内标注文后参考文献,图表下方注释文后参考文献,括号注明 URL 地址,图表下注释来源,页脚注释来源,正文文字说明来源,文后致谢来源,无引用标注

由于数据引用目前还没有规范统一的格式及可靠的自动识别方法,本研究采用人工识别方式收集样本论文的科学数据引用数据。阅读全文实施编码步骤如下:以每篇论文中涉及科学数据的句子、词语、呈现形式及相关上下文为分析单元,重点关注文中的图表数据与文本描述中出现的定量数据;识别科学数据引用记录,重点区分科学数据引用与文献引用、事实陈述、科学数据提供、数字描述(如日期、自然数)等情况,根据内容分析类目系统的编码体系对每条引用纪录内容进行多维度编码,结果存储为 Excel 文件;清洗、整理 Excel 文件后,采用频数统计、交叉列联表等分析不同维度科学数据引用特征。

(2)对应分析法。对应分析法也称关联分析法,是一种多元相依变量统计分析技术。它通过分析定性变量构成的交互汇总表来揭示变量间的联系,能够将两组看不出明确联系的数据通过视觉上可以接受的定位图展现出来,以此揭示同一变量各个类别之间的差异,以及不同变量各个类别之间的对应关系[16]。

基于量化后的编码数据文件,采用 IBM SPSS Statistic 的对应分析模块对不同维度的科学数据引用特征变量进行关联分析,通过不同变量及其元素间的对应关系探索科学数据的引用模式特征。

3 研究结果

3.1 科学数据引用特征

通过人工识别对 175 篇样本论文进行初步统计编码、数据两次校验与量化处理,最终整理出 632 条科学数据引用记录,应用描述统计方法分别从科学数据引用量的总体分布和 6 个不同维度的具体分布描述科学数据引用特征。

3.1.1 科学数据引用量总体分布 6 种期刊 2017 与 2018 年第一期刊载论文的科学数据引用情况如表 3、表 4 所示。总体上,2017 年第一期的 89 篇论文中有科学数据引用的论文为 70 篇,篇均引用量为 4.64 条;

表 3 2017 年第一期各期刊科学数据引用分布

期刊	论文总数(篇)	有引用论文(篇)	无引用论文(篇)	科学数据引用量(条)	平均引用量(条/篇)
中国图书馆学报	7	5	2	29	5.80
大学图书馆学报	17	15	2	77	5.13
图书与情报	21	13	8	48	3.69
图书情报知识	15	13	2	46	3.54
国家图书馆学刊	13	9	4	43	4.78
图书情报工作	16	15	1	82	5.47
合计	89	70	19	325	4.64

表 4　2018 年第一期各期刊科学数据引用分布

期刊	论文总数（篇）	有引用论文（篇）	无引用论文（篇）	科学数据引用量（条）	平均引用量（条/篇）
中国图书馆学报	8	8	0	33	4.13
大学图书馆学报	18	15	3	96	6.40
图书与情报	18	16	2	63	3.94
图书情报知识	13	6	7	44	7.33
国家图书馆学刊	11	7	4	22	3.14
图书情报工作	18	13	5	49	3.77
合计	86	65	21	307	4.72

2018 年第一期的 86 篇论文中有科学数据引用的论文为 65 篇，篇均引用量为 4.72 条。以上数据表明，图书情报期刊论文的科学数据引用较为普遍，近两年的科学数据引用量较为稳定。

3.1.2　科学数据类型分布　科学数据类型的分布（见图 1）表明：我国图书情报领域科研工作者对统计整理类科学数据的引用量最大，约占总引用量 78%，其次分别为对研究成果类（占比约 17%）和原始记录类（4%）。统计整理类科学数据主要包括来自各行各业各种用途的调查统计数据，是经过人工处理与整理的次生性科学数据；研究成果类科学数据指科研活动中产生的非统计类科学数据，主要包括科学研究产生的公式、算法、模型、量表、经验数据等；原始记录类科学数据指未经人工统计与整理工作形成的记录型科学数据，是呈现研究对象运动的原始记录，样本论文中主要涉及对互联网平台网页生成的原始数据（网页点击量、日期）的引用。

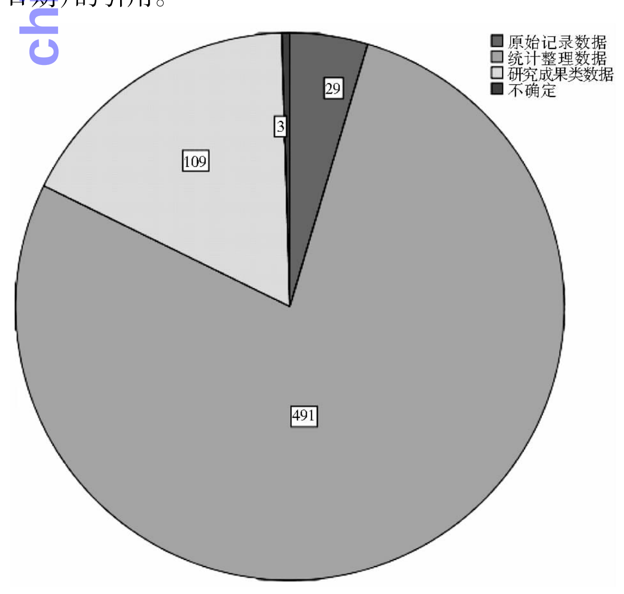


图 1　科学数据类型分布

3.1.3　科学数据来源地区分布　从科学数据的来源地区分布（见图 2）可以看出：我国图书情报科研工作者对国内（占比约 47.47%）和国外科学数据（41.14%）的引用量没有太大差异，表明本领域的科研活动广泛需要来自世界不同国家的科学数据，国内外科学数据体现出同等重要性。此外，不能判断科学数据来源的数据占比 11.39%，原因主要是文中没有对科学数据的来源进行引用标注，且引用上下文内容也难以判断该科学数据的来源。

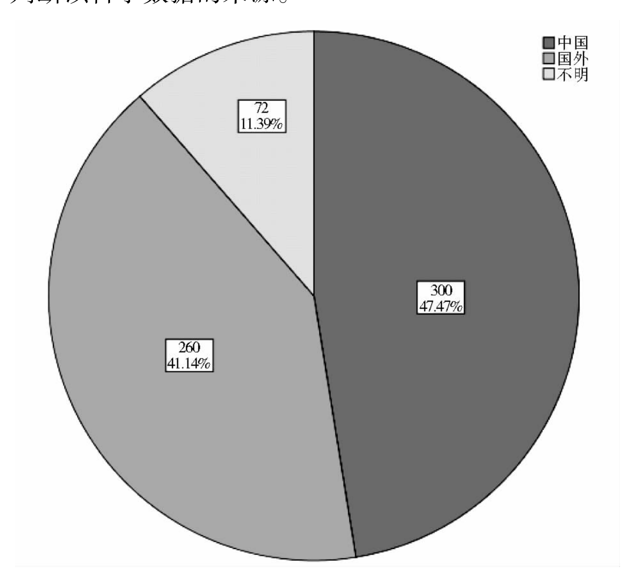


图 2　科学数据来源地区分布

3.1.4　科学数据创建者分布　科学数据创建者指科学数据形成的初始来源。统计结果（见图 3）表明：引用个人研究形成的科学数据最多（占比约 28.96%）；第二来源是社会团体的科学数据（11.55%），该类数据是指非营利性的学术型、行业型、专业性、联合性社会团体形成的科学数据；第三来源是商业型企业的科学数据（7.44%）；此外依次为来自学校图书馆（7.12%）、数据库（6.65%）、政府行政机关（6.17%）、公共图书馆（5.22%）、学校（3.96%）、科研机构（0.63%）以及档案馆（0.47%）的科学数据。另外，约 16.46% 的科学数据引用记录无法判断科学数据创建来源，说明目前图书情报领域科研工作者对科学数据创建来源的引用标识意识有待提高，无引用行为标注较为明显。

3.1.5　科学数据传播者分布　科学数据引用传播者指科学数据的引用来源，需要与科学数据引用创建者相区分。科学数据创建者指科学数据形成的原始来源，表示“该科学数据在哪里形成”，而科学数据引用传播者指科学数据的引用来源，表示“从哪里引用的该科学数据”。

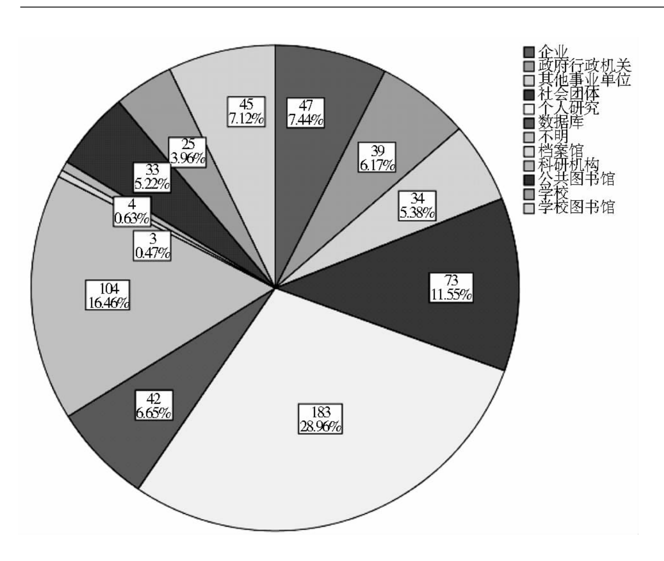


图 3 科学数据创建者类型分布

科学数据的传播者类型分布(见图 4)表明:我国图书情报领域的科学数据引用渠道较为多样,期刊论文作为数据传播渠道受到科研工作者的青睐(约占比 25.32%);其次包括社会团体(9.65%)、企业(8.86%)、教育机构(8.23%)等机构部门提供的互联网平台。值得关注的是,存在大量无法识别科学数据引用渠道的情况,结合科学数据创建来源的分布情况可以看出,我国图书情报领域科研工作者对科学数据创建来源和传播渠道来源的标注意识略有欠缺,传播渠道来源的无引用标注情况更为突出。

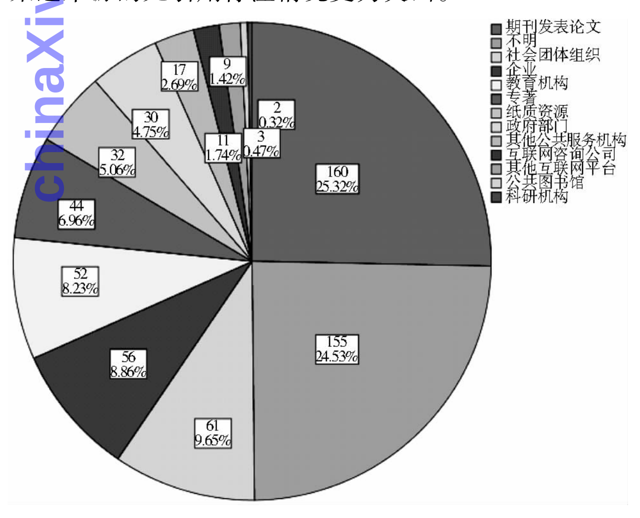


图 4 科学数据传播者类型分布

3.1.6 科学数据引用呈现形式分布 科学数据引用呈现类型是科学数据引用行为的直接表现,指科学数据引用在论文中的具体呈现形式,如文字论述形式、图表方式、公式和模型等。

科学数据引用呈现类型分布(见图 5)表明:超过 80% 的科学数据引用记录是以文字形式呈现引用科学数据,说明文字形式是我国图书情报领域科研工作者的偏好引用形式;其次,基于引用数据作表(11.55%)与直接引用图(6.49%)的形式也较为常见。

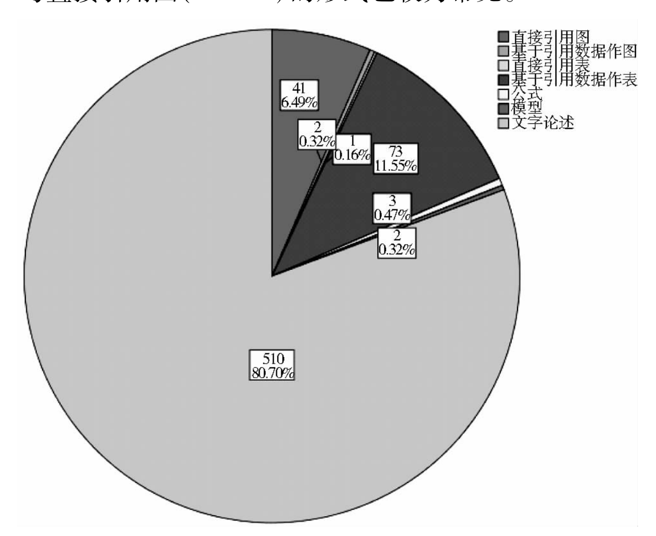


图 5 科学数据引用呈现类型分布

3.1.7 科学数据引用标注方式分布 科学数据引用标注方式分布(见图 6)表明:科学数据引用记录的引用标注方式较多样化,标注文后参考文献是最为常见的科学数据引用标注方式,约占 60%。除此之外的其他各种引用标注方式包括:图表名称标注参考文献、图表内标注参考文献、注明 URL 地址、图表下注释来源、页脚注释来源、正文文字说明引用来源、文后致谢引用来源等,但占比均较少;此外,无引用标注情况较为明显,约占 22%。

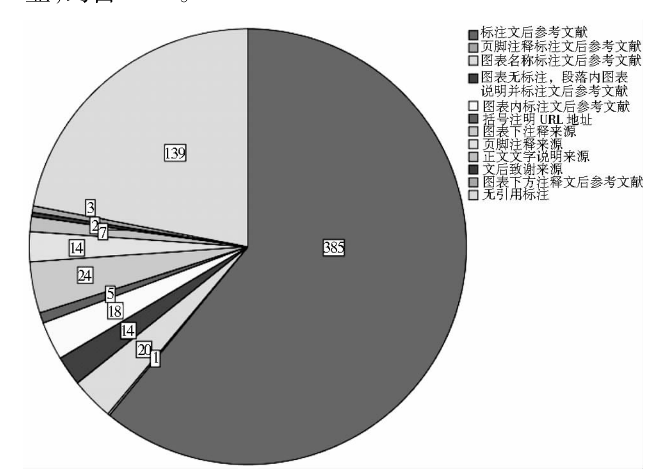


图 6 科学数据引用标注方式分布

3.2 科学数据引用特征关联性

描述统计数据主要从科学数据内容属性和引用行为方式两方面揭示了不同维度的科学数据引用分布特征。那么,这些特征间是否存在关联?不同期刊对不同来源科学数据的引用是否存在一定倾向?科学数据

类型与科研工作者的引用呈现方式、引用标注方式之间是否存在关联？不同创建来源的科学数据与科研工作者选择的引用渠道之间具有怎样的关联？

基于以上问题，本研究以科学数据引用特征相关变量之间的关系提出 5 个研究假设，并采用对应分析方法对假设进行检验和分析。经 SPSS 对变量数据的分析运算，5 个研究假设内容及其对应分析重要运算结果数据如表 5 所示：

表 5　科学数据引用特征关联假设与对应分析结果

序号	研究假设	卡方检验伴随概率 P 值	验证结果
H1	期刊种类与科学数据创建者类型相关	0.000	通过
H2	科学数据类型与科学数据引用呈现形式相关	0.047	通过
H3	科学数据类型与科学数据引用标注方式相关	0.000	通过
H4	科学数据引用呈现方式与引用标注方式相关	0.000	通过
H5	科学数据创建者类型与科学数据传播者类型相关	0.000	通过

根据对应分析计算的卡方检验数据，5 个假设的伴随概率 P 值均小于 0.05，表明假设均通过检验，相关变量之间具有较高的相关度。同时，对应分析经降维处理，以行、列变量的各类别元素的得分情况绘制对应关系的二维散点图，其分布情况能够解释变量间潜在的相关关系，即不同维度变量的远近距离体现其亲疏关系，距离越近的变量元素对应关系越密切。由此，以下结合各假设的对应分析结果探索不同维度科学数据引用特征间的关联关系，以揭示引用行为的内在规律。

3.2.1　期刊种类与科学数据创建者类型关联分析　图 7 是 6 个类别的期刊变量和 13 个类别的科学数据创建者的对应分析结果。以横坐标（维 1）与纵坐标（维 2）的原点为界划分 4 个象限，位于不同象限的变量距离较远，表明之间对应关系较弱，反之位于相同区域的变量间对应关系较强，且距离越近的关系越紧密。总体来看，期刊变量点与科学数据创建者变量点分布在不同象限且不同变量元素之间的距离远近不一，表明期刊与科学数据创建来源存在一定的对应关系及引用倾向性。具体而言，《图书与情报》与政府行政机关、社会团体、科研机构科学数据均位于第 1 象限，说明该期刊的刊载论文主要引用以上三种科学数据，同时期刊变量点与政府行政机关、社会团体的距离较近，说明其较多引用由政府行政机关和社会团体创建的科学数据，而对科研机构科学数据的引用较少；同理观察其它变量分布可以看出，《图书情报工作》对个人研究形成的科学数据引用较密集，同时也引用少量来自数据库和档案馆的科学数据；《大学图书馆学报》主要引用公共图书馆、学校创建的科学数据；《国家图书馆学刊》主要引用学校图书馆、公共图书馆形成的科学数据；《中国图书馆学报》与《图书情报知识》的引用情况较为相似，主要引用来自企业、其他公共服务机构的科学数据。根据图示分布情况推测，不同期刊的引用倾向性可能与期刊性质、主办者、选稿主题及刊载论文内容相关。

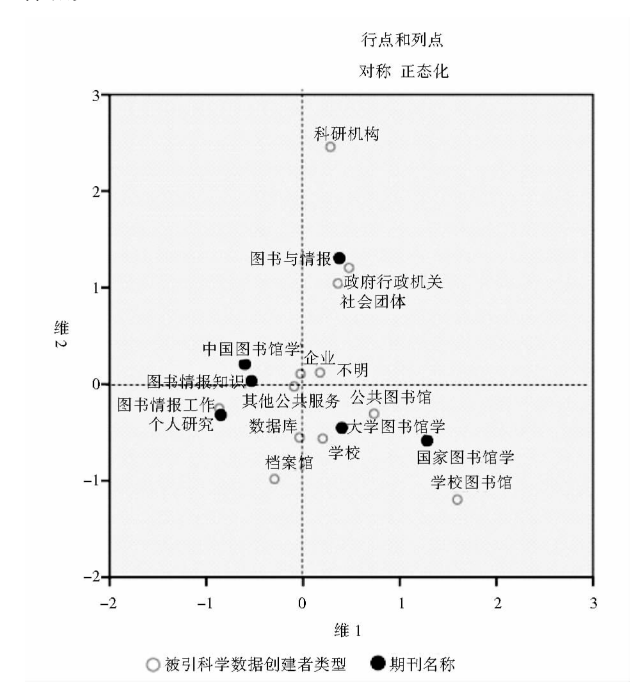


图 7　期刊与科学数据创建者类型对应分析结果

3.2.2　科学数据类型与科学数据引用呈现形式关联分析　结合变量点分布的区域位置和距离关系观察 H2 对应分析结果图像（见图 8）可以看出，科学数据类型与其被引用的呈现形式之间存在一定对应关系。具体而言，第 1 象限中，围绕统计整理类科学数据的引用呈现形式变量点较为密集，表明科研工作者引用该类科学数据的方式较多，主要倾向以文字论述形式进行引用，其次包括基于引用数据作表、基于引用数据作图或直接引用表格；第 2 象限显示原始记录类科学数据与直接引用图方式关系紧密，表明科研工作者对原始记录类科学数据倾向为直接引用图；第 3 象限显示引用研究成果类科学数据对应的引用呈现形式主要为公式、模型形式；此外，不确定类型的科学数据主要由不规范引用和无引用造成，因此与任何变量点的相关性较弱。

3.2.3　科学数据类型与科学数据引用标注方式关联分析　观察 H3 对应分析结果（见图 9）可以看出，三种

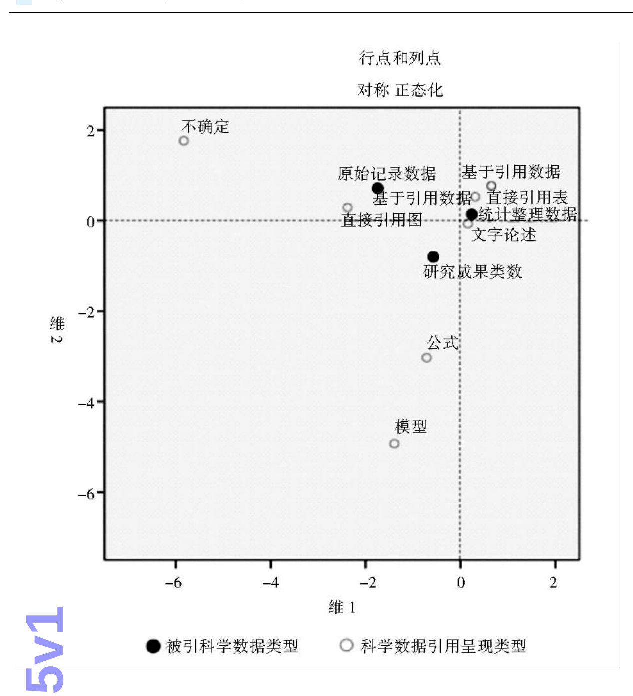


图 8 科学数据类型和科学数据引用呈现形式对应分析结果

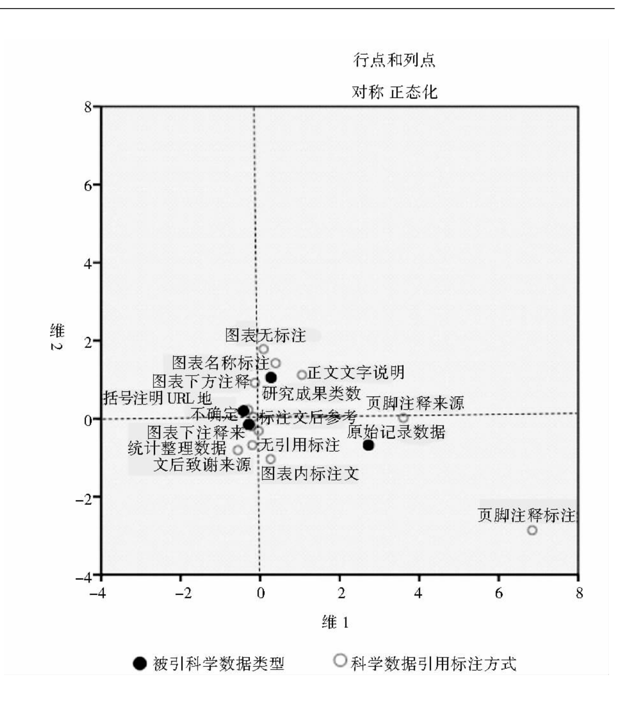


图 9 科学数据类型与科学数据引用标注方式对应分析结果

不同类型科学数据变量点分别位于图像的 1、3、4 象限，说明不同类型科学数据与被引用标注方式的对应关系存在明显差异。第 1 象限变量对应关系显示，研究成果类科学数据对应密切的引用标注方式分别为：图表下方注释参考文献、图表名称标注参考文献、正文文字说明来源、段落正文说明来源并标注参考文献；第 3 象限显示，围绕统计整理类科学数据的引用标注方式较为密集，其对应密切程度由强至弱分别为：标注文后参考文献、无引用标注、括号标注来源 URL 地址、图表下注释来源、文后致谢来源；第 4 象限显示，对应原始记录类科学数据的引用标注方式主要为页脚注释来源，其次为图表内标注参考文献、页脚注释参考文献。

3.2.4 科学数据引用呈现方式与引用标注方式关联分析 H4 对应分析结果（见图 10）显示，科学数据的 8 种引用呈现形式与 11 种不同引用标注方式之间存在一定对应关联。第 1 象限变量对应关系显示，引用科学数据作表格时主要倾向使用的引用标注方式为图表下注释来源、图表内标注参考文献及文后致谢来源；文字论述、模型、公式三种引用呈现形式的变量点在第 2、3 象限交界处交叉重叠，说明三者对应的引用标注方式基本没有差异，主要包括标注文后参考文献和括号标注 URL 地址；第 4 象限中存在直接引用图像科学数据、表格科学数据和引用科学数据作图三种变量点，说明三者对应的引用标注方式较为相似，根据变量分布对应距离具体细分来看，引用图像科学数据时主要倾向的引用标注方式为正文文字说明来源、图表名称标注参考文献、页脚注释来源、段落正文说明来源并标注参考文献，而引用表格科学数据时主要倾向使用在页脚注释并标注参考文献，引用科学数据作图时主要倾向于使用在图表下方注释并标注参考文献。

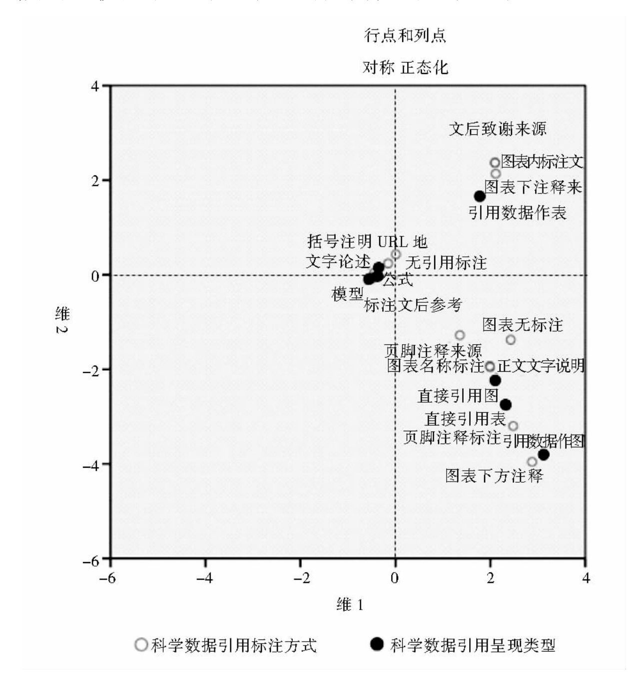


图 10 科学数据引用呈现方式与科学数据引用标注方式对应分析结果

3.2.5 科学数据创建者类型与科学数据传播者类型关联分析 科学数据引用强调对被引科学数据的来源

信息进行标注与说明，对应《信息技术 科学数据引用》国家标准，其中体现科学数据来源出处的标注元素包括科学数据的创建者和传播者，二者分别从原始创建和传播渠道两方面定位科学数据来源。从科学数据来源的引用一致性出发，科学创建者类型与传播者类型应当基本对应一致。

假设 5 通过两者变量的对应关系验证其对应一致情况，结果见图 11。总体上看，12 种科学数据创建者类型与 10 种传播者类型呈现较为一致的对应关系，如第 1 象限“政府部门”与“政府行政机关”变量点距离较近，表明政府部门创建的科学数据基本由政府部门数据公开平台引用；第 2 象限“个人研究”与“期刊论文”变量点基本重叠，表明个人研究形成的科学数据基本从期刊论文中引用。然而，图中显示不同类别变量的对应距离存在差距，说明科学数据创建者类型与传播者类型的引用一致性在不同情况下存在差异，由此通过交叉列联分析进一步探究二者的对应一致性情况。

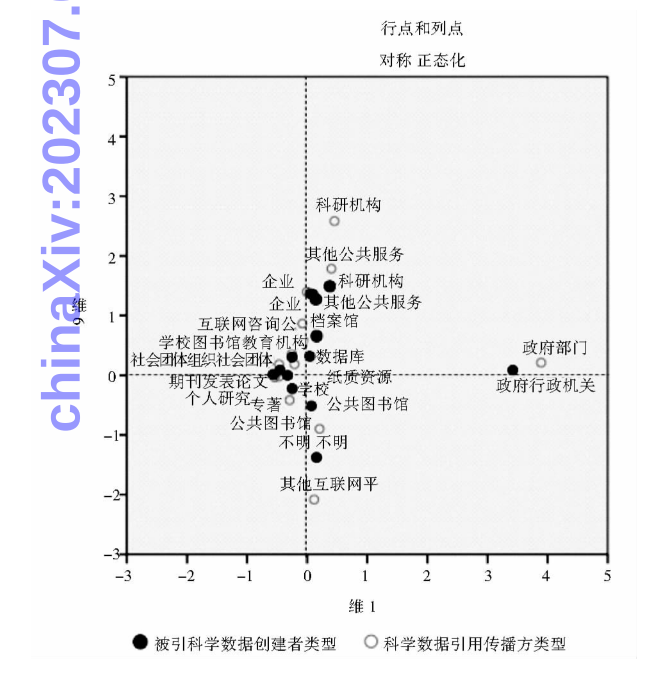


图 11　科学数据创建者类型与科学数据传播者类型对应分析结果

不同类型科学数据创建者与其对应的引用来源（科学数据传播者类型）一致性结果数据显示（见表 6）：632 条科学数据引用记录中，423 条科学数据引用为创建者与传播者一致，而 41 条为不一致，168 条不能确定是否一致。总体上科学数据创建者类型与科学数据传播者类型的引用一致率为 66.93%，表明多数情况科研工作者是从科学数据的创建之处进行直接引用，无法确定科学数据创建来源和引用来源一致性的情况占总比 26.58%，这主要与科学数据不规范引用和无引用行为相关。对比不同创建来源科学数据的引用一致率发现，引用一致性较高的是来源个人研究和学校图书馆的科学数据，其一致率分别达 93.99% 与 93.3%，说明我国图书情报领域科研工作者对这两类来源的科学数据基本上从原始创建处获取引用；来源为政府行政机关、公共部门、社会团体、公共图书馆的科学数据引用一致率在 70% －80% 之间，来源为企业、档案馆、学校的科学数据引用一致率在 60% －70% 之间，说明相关部门、机构通过互联网平台公开发布的科学数据在传播中可能容易造成科学数据源地址的模糊，从而导致科学数据的引用渠道并非其创建者的原始发布来源，造成“二手科学数据引用”现象；来源为数据库的科学数据引用一致率为 54.76%，而来源为科研机构的引用一致率仅为 25%，其引用一致率较低主要由无引用行为造成，说明科研工作者对数据库和科研机构来源科学数据的引用意识有待重点加强。

表 6　被引科学数据创建者与传播者一致性交叉列联表

被引科学数据创建者类型	创建者与传播者一致性			总计	引用一致率
	一致	不一致	不确定		
个人研究	172	6	5	183	93.99%
学校图书馆	42	3	0	45	93.33%
政府行政机关	31	2	6	39	79.49%
社会团体	56	7	10	73	76.71%
其他公共服务部门	26	5	3	34	76.47%
公共图书馆	24	8	1	33	72.73%
学校	17	4	4	25	68.00%
档案馆	2	0	1	3	66.67%
企业	29	4	14	47	61.70%
数据库	23	2	17	42	54.76%
科研机构	1	0	3	4	25.00%
不明	－	－	104	104	－
总计	423	41	168	632	66.93%

4　讨论与结论

本研究通过调研我国 6 种图书情报领域期刊 175 篇论文揭示该领域科学数据引用特征与引用特征之间的对应关系，主要获得以下结论：

（1）图书情报领域期刊论文对科学数据的引用较为普遍。引用科学数据的论文占比 77.14%，2017 年第一期与 2018 年第一期的科学数据引用量较为稳定，篇均引用量分别为 4.64 条与 4.72 条。以往研究表明，《中国图书馆学报》《情报学报》和《大学图书馆学

报》2003－2013年刊载论文中引用科学数据的论文占比49.89%，在2003～2008年间篇均引用次数均在1～2次之间，2009年开始逐年提升并在2013年达到2.7次[4]。对比本研究数据结果可以发现，近年我国图书情报领域期刊论文中有科学数据引用的论文占比量和科学数据篇均引用量均明显提高，在一定程度上说明近年来图书情报领域以数据为基础的实证研究正逐渐走强，体现了数据密集科学范式的发展趋势；而从横向比较来看，社会学领域2003－2014年刊载论文的篇均数据引用次数超过6次[5]，明显高于图书情报领域的当前引用水平，一方面说明社会学领域科学研究对数据的依赖程度较高，另一方面表明图书情报领域在数据使用方面还有较大提升空间。

（2）图书情报领域科研工作者对统计整理类科学数据的引用最为普遍，对研究成果类和原始记录类科学数据的引用较少。这一结果与过去调研的社会科学领域期刊论文中的引用情况较为一致，如美国校际社会科学数据共享联盟存储库（ICPSR）的论文中对调查类数据的引用量最大[1]，在我国社会学期刊论文中最常见引用的科学数据是人口普查资料、年鉴资料等正式出版的宏观调查统计数据[5]。科学数据类型与引用行为方式的关联分析发现，一是统计整理类数据主要通过文字论述直接引用，另外存在部分基于引用数据绘制表格和图像的情况，而原始记录类和研究成果类数据的引用主要表现为直接引用图像、公式、模型等；二是科学数据的引用标注方式较为多样，缺乏统一性，科研工作者不仅对不同类型科学数据的引用标注方式具有差异，而且对相似类型科学数据的引用标注方式也不尽相同。总体上，传统文献引用方式为科学数据引用标注的通用方式，尤其对于引用量最大的统计整理类科学数据而言，“标注文后参考文献”是最为普遍的引用标注方式，而关于图像、表格、公式等原始记录类和研究成果类科学数据的引用标注方式较为多样化，以页脚、文内、图表名称、图表下方、参考文献等不同位置进行不同形式的标注广泛存在。

（3）图书情报领域的科学数据引用来源范围广泛。地区来源方面，对国内与国外科学数据的引用量较为均衡；创建来源方面，广泛引用个人研究、社会团体、企业、政府部门等众多来源的科学数据，其中对个人研究科学数据的引用最为普遍。关联分析发现，科学数据的不同创建来源对应不同类型的引用渠道，引用量最大的个人研究科学数据主要从期刊论文中获取，相关机构或团体创建的科学数据主要由官方互联网平台获取。由于科学数据是在专门的科研活动中形成，于创建者而言具有归属性，从原始创建处引用科学数据并标注来源应是规范的引用行为，即科学数据创建来源和引用来源应具有一致性，然而分析二者变量对应关系发现，虽然多数情况两者具有一致性，但仍然存在不同程度的不一致现象，即存在一定数量的“二手科学数据引用”。通过进一步对比发现，针对不同来源科学数据的二手引用程度具有较大差异，引用来自个人研究和学校图书馆的科学数据时二手引用较少，而引用来自公共图书馆、社会团体等机构科学数据时较多。有学者提出科学数据创建来源的开放共享是影响科学数据引用的关键因素[1]，由此可推测科学数据二手引用现象可能与科学数据发布、传播渠道的开放程度有关：个人研究形成的科学数据通常包含在发表的论文或互联网开放平台中，我国的论文数据库例如CNKI、万方等开放程度较高，因此对个人研究的科学数据易获取性较强，不易形成二手引用行为；同理，社会团体、公共服务部门、政府机关在一定情况下会公开发布调查研究所得的相关数据，大数据环境下互联网平台为各个来源的科学数据传播提供了新渠道，提高了科学数据的多渠道获取和易获取性；而档案馆、科研机构的开放程度较低，其科学数据的易获取性较弱，较容易导致二手引用行为。

（4）图书情报领域科研工作者科学数据的无引用行为较为突出。不同维度的引用特征统计数据表明，无法通过论文内引用标注或引用段落上下文内容识别该维度分类的情况约占22%－26%，一方面表明科研工作者对科学数据引用没有形成与参与文献引用相似的意识理念，另一方面也表明没有成熟的引用标准规范来引导大家的科学数据引用行为。

本研究的相关结论为科学数据引用的相关管理工作提供了一些启示：

（1）科学数据的引用分布在一定程度上反应了一个领域科学数据结构与引用需求，其结果为规划领域科学数据仓储、开发引用渠道提供了参考：一方面，对伴随学术论文发表时形成的个人研究科学数据进行统一管理，根据情况提供数据的共享渠道、规范数据引用形式；另外，广泛收集国内外重要机构和社会团体发布的科学数据，建立统一科学数据获取平台或提供引用链接渠道。通过规范科学数据获取平台和引用渠道有利于提高科学数据的可见性和开放共享程度，从而减少科学数据的二手引用行为。

（2）科学数据引用实践需要统一通用的科学数据

引用规范进行指导。本研究发现科学数据引用标准在图书情报领域未得到有效实施,大部分作者仍然参考的是传统文献引用方式,其它引用标注方式众多,缺乏统一性,其原因可能是科学数据引用规范不如传统文献引用深入人心,也可能是引用规范在实际中不能完全满足各种类型科学数据的引用要求,亦或是科研工作者的引用意识欠缺所致。根据研究结果推测,参考文献方式可能较符合目前科研工作者的引用标注习惯,但该方式仅能识别科学数据的引用来源信息,无法指示科学数据的内容、创建者、形成时间等其它信息,如若科学数据的引用标注能与传统的文献引用相结合,则需要根据科学数据类型差异对参考文献的标注元素进行补充完善,使之适合科学数据的引用要求;此外,论文中图表类型科学数据的引用标注方式和标注位置较为混乱,需要针对不同类型科学数据给出明确的引用标注位置和方式。

最后,本研究尚存在一定局限:仅以 CNKI 数据库收录的"图书情报与数字图书馆领域"六种期刊 175 篇论文为样本,样本量较为有限,结论的普适性有待通过更多领域、更多数据来进一步证明;仅以科学数据引用结果记录分析科学数据引用行为特征与内在规律,对于各种现象形成的深层原因、影响科研工作者引用行为的影响因素没有涉及;基于人工识别编码的内容分析虽然可以识别出一些现象和规律,但研究过程效率较低、工作繁琐,需要开发科学数据引用的自动识别工具。

参考文献:

[1] 屈亚杰,王亚男. 社会科学领域科学数据的引用现状与特点分析[J]. 数字图书馆论坛,2017(6):25－31.
[2] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. 中华人民共和国国家标准 GB/T 35294－2017《信息技术 科学数据引用》[EB/OL]. [2018－08－20]. http://www.gb688.cn/bzgk/gb/newGbInfo? hcno = A495CA355BAF00D962AA8DD84C3B2C16.
[3] 王雪,马胜利,佘曾溧,等. 科学数据的引用行为及其影响力研究[J]. 情报学报,2016,35(11):1132－1139.
[4] 丁楠,丁莹,杨柳,等. 我国图书情报领域数据引用行为分析[J]. 中国图书馆学报,2014(6):105－114.
[5] 丁楠,杨柳,丁莹,等. 我国社会学期刊论文数据引用行为研究[J]. 图书与情报,2014(6):88－93.
[6] WILLIAMS S C. Data practices in the crop sciences: a review of selected faculty publications[J]. Journal of agricultural & food Information,2012,13(4):308－325.
[7] PARK H, YOU S, WOLFRAM D. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields[J]. Journal of the Association for Information Science and Technology, 2018, 69(11): 1346－1354.
[8] 赵蕊菡. 科学数据论文的重用现状研究——基于数据期刊"Earth System Science Data"的引文分析[J]. 情报理论与实践, 2017, 40(11): 52－57, 72.
[9] 孟祥保, 钱鹏. 数据生命周期视角下人文社会科学数据特征研究[J]. 图书情报知识, 2017(1): 76－88.
[10] CHEN R S, DOWNS R R, SCHUMACHE R J A. Analyzing data citations to assess the scientific and societal value of scientific data[EB/OL]. [2018－09－26]. http://academiccommons.Columbia.edu/catalog/ac:157135.
[11] HENDERSON T, KOTZ D. Data citation practices in the CRAWDAD wireless network data archive[C]// Chengdu: IEEE international conference on communication software & networks. 2015.
[12] PILAT D, FUKASAKU Y. OECD principles and guidelines for access to research data from public funding[J]. Data science journal,2007,6:4－11.
[13] 李慧佳,马建玲,王楠,等. 国内外科学数据的组织与管理研究进展[J]. 图书情报工作,2013,57(23): 130－136.
[14] 李丹丹,吴振新. 研究数据引用研究[J]. 图书馆杂志, 2013, 32(5): 65－71.
[15] 黄国彬,刘馨然,姜颖. 影响科学数据引用的外部因素分析[J]. 数字图书馆论坛, 2017(6): 2－8.
[16] 杜强,贾丽艳. SPSS 统计分析从入门到精通[M]. 北京:人民邮件出版社, 2009: 419－422.

作者贡献说明:

丁文姚:数据搜集、整理与分析,撰写论文初稿;
李健:参与论文修改;
韩毅:整体研究设计与规划,论文修改完善。

Research on the Characteristics of Scientific Data Citation in Journal Articles in Library and Information Science in China

Ding Wenyao　Li Jian　Han Yi
College of Computer and Information Science, Southwest University, Chongqing 400715

Abstract: [**Purpose/significance**] Exploring the characteristics and laws of the scientific data citation behavior of journal articles not only helps to describe the utilization of scientific data in academic areas, but also reveals the data citation models of the expression of academic works. [**Method/process**] The articles contained in the first issue of six jour-

nals in library and information science in China during 2017 and 2018 were selected as the sample. Based on the citation elements of the China's national standard Information Technology Scientific Data Citation, the scientific data citation behavior records of the sample papers had been encoded from nine dimensions by the content analysis method, and the characteristics and the relationships between different characteristics of the citation behavior had been analyzed by the statistical analysis method. [**Result/conclusion**] The results show that: scientific data citation is relatively common in library and information science, and the reorganized statistical data from the domestic and international are widely cited, and the citation of personal research scientific data in journal academic articles is large; there are some correspondence relationships between citation styles and scientific data types, but diverse citation styles lack uniformity; in addition, second-hand citation behavior is obvious, and its degree is related to the creator type of scientific data.

Keywords: library and information science journal articles scientific data citation citation characteristic

第五届中国特色新型智库建设学术研讨会在西安举办

智库是党和政府科学、民主决策的智力支撑。如何贯彻习近平新时代中国特色社会主义思想,发挥智库为重大决策咨询服务的作用,做好资政建言,是当前中国智库面临的重要议题,也是全面深化改革的重要举措。2019 年 11 月 10 日,以“数据驱动的新型智库建设”为主题的第五届中国特色新型智库建设学术研讨会在西安市唐城宾馆顺利召开。

本次会议由中国科学院文献情报中心与西安电子科技大学联合主办,由中国科学院文献情报中心《智库理论与实践》编辑部、西安电子科技大学科学数据管理与区域政策研究中心、陕西信息资源研究中心、西安电子科技大学经济与管理学院联合承办,同方知网(北京)技术有限公司协办。来自国内研究所、高校、企业等 50 多家单位的 150 多位专家学者参加了会议。

中国科学院文献情报中心副主任、《智库理论与实践》主编刘细文研究员和西安电子科技大学总会计师谢军占出席会议并致辞,开幕式由中国科学院大学图书情报与档案管理系主任、《智库理论与实践》执行副主编初景利教授主持。

大会主题报告分为上午和下午两个阶段。在上午的报告议程中,南京大学的李刚教授、北京大学王继民教授、陕西省科学技术情报研究院张薇院长、刘细文研究员、中国知网党政与金融知识管理事业本部谢磊总经理围绕有关智库建设与发展等方面内容进行了精彩汇报。在下午的报告议程中,北京拓尔思信息技术股份有限公司施水才总裁、重庆大学图书馆杨新涯馆长、中国人民大学朝乐门副教授、西安电子科技大学赵捧未教授做了精彩报告。

初景利教授做大会总结。他强调了数据对智库的作用,简明分析了图情档学科与智库研究的关系,提出了从情报到智库的“最后一公里”问题。最后,初景利教授预告了《智库理论与实践》2020 年将发起举办的会议:2020 年 7 月于济南举办第五届智库能力与新型智库建设高级研修班,主题为“新型智库建设的理论与方法”;2020 年 10 月在徐州举办第六届中国特色新型智库建设学术研讨会,主题为“地方智库发展模式与能力建设”。

本次大会报告旁征博引、深入浅出、内容丰富、信息量大,充分体现了各位专家学者对中国特色新型智库建设的深刻洞察力,为学者们了解智库领域的最新动态、最新研究成果和探讨新型智库建设提供了一个良好的交流平台。

《智库理论与实践》编辑部供稿